Domestic Travelling and Tourism Post the National (India) Lockdown

A data driven guide

- By Chris Pinto

01

Agenda

02 —

Introduction
Data
Methodology
Results
Discussion
Conclusion

Introduction

03 —

1. THE PANDEMIC

With coronavirus causing chaos in our everyday lives, what does it mean if you're an avid traveller? What does it mean to those that love to travel but now have to worry about safety?

2. THE FUTURE OF TRAVEL

It's unlikely that peope will refrain from travel. As said by Andrew McCarthy, 'The farther I travel, the closer I am to myself. But with international travel looking hazy, domestic is the new international. It's funny that personally, I have travelled more countries than states in my own country. I know there are many more like me.

3. AIM OF THE PROJECT

Keeping the current scenario in mind, the aim of the project is to find and cluster touristic hotspots and destinations based on varying factors like popularity, safety (in terms of coronavirus cases) and similarity. This would serve as a guide to people as to where they could make that travel plan they so wish to do.

Data

What are the sources and types of data?

04 —

1. Google Travel for web scraping touristic destinations.
https://www.google.com/travel/

2. Google Places, Geocoding and Reverse Geocoding APIs due to their robustness in India.
https://developers.google.com/places/web-service/intro

3. Wikipedia for web scraping information about states and districts, particulary their populations and area.
https://en.wikipedia.org/wiki/List_of_districts_in_India

4. Publicly available coronavirus APIs to gather active cases in states and districts.
https://api.covid19india.org/

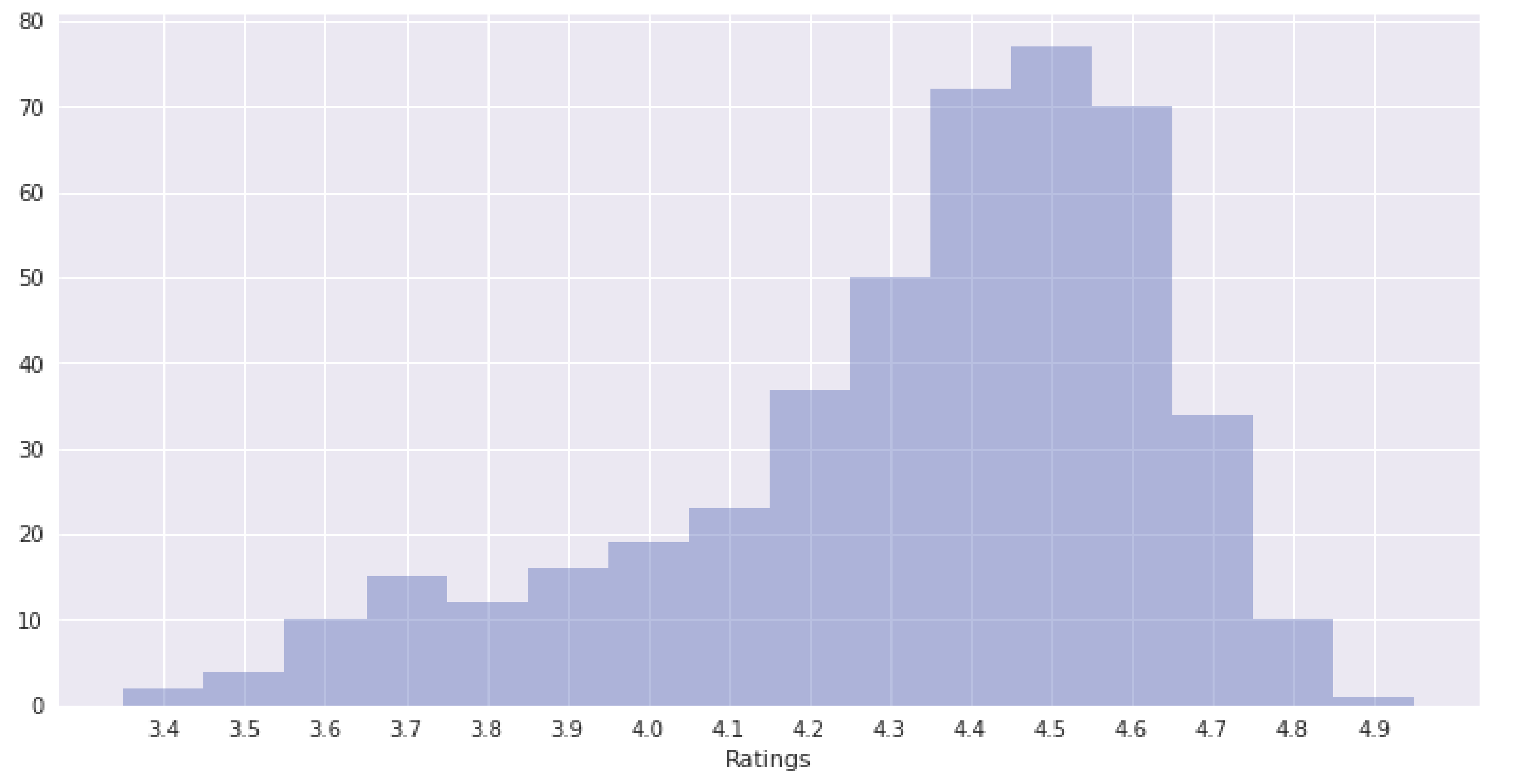
→

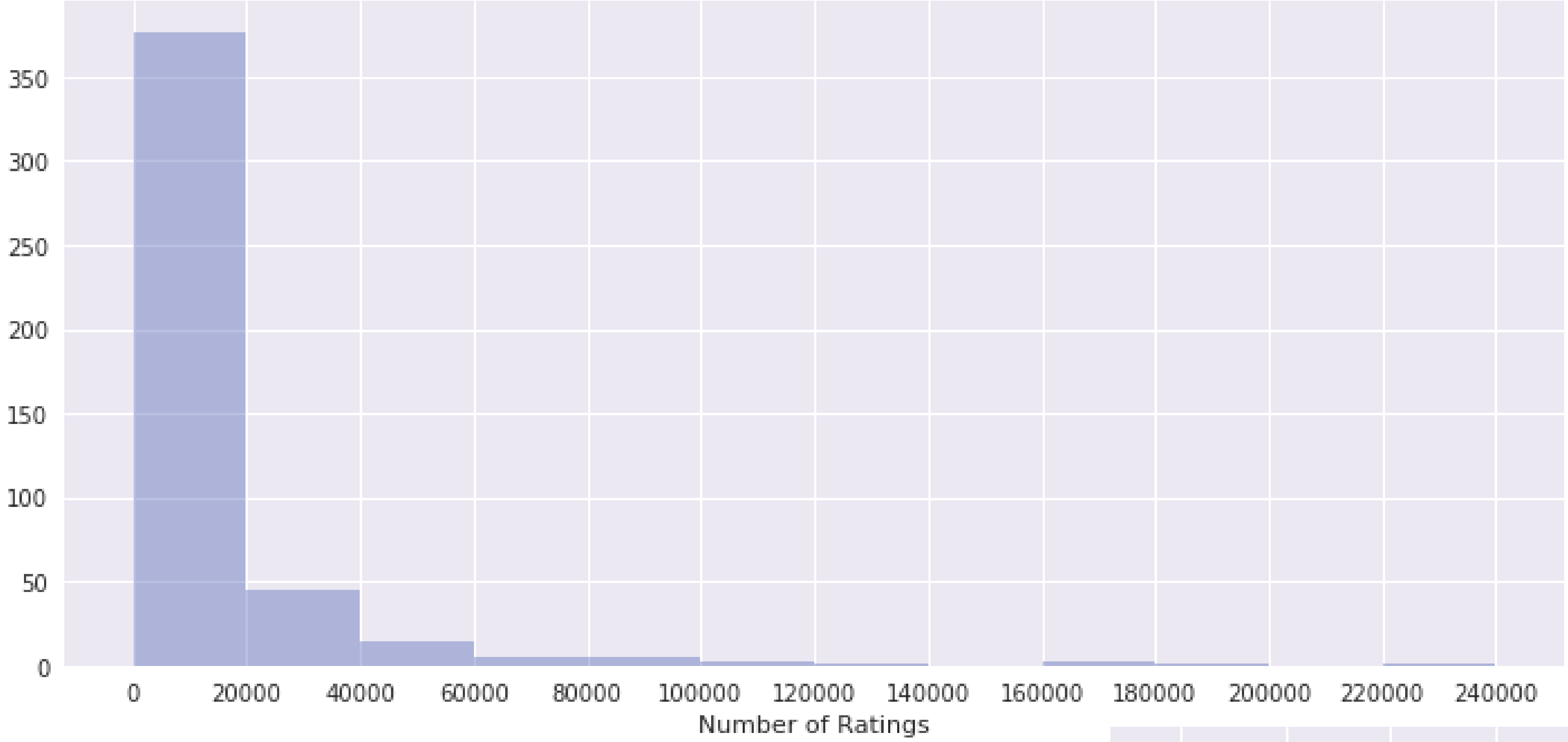
Methodology

Exploratory analysis and reporting

05

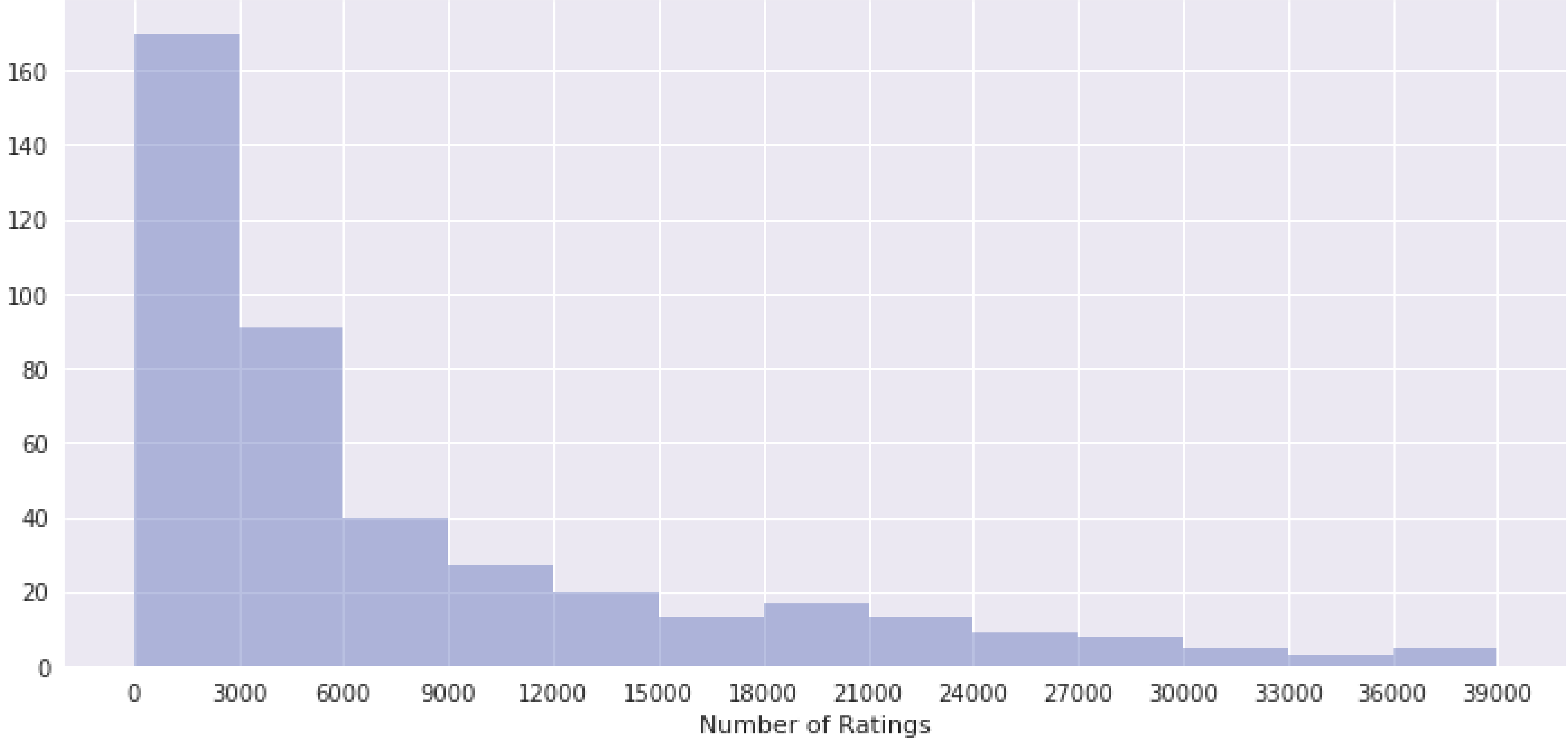
Histogram of Rating

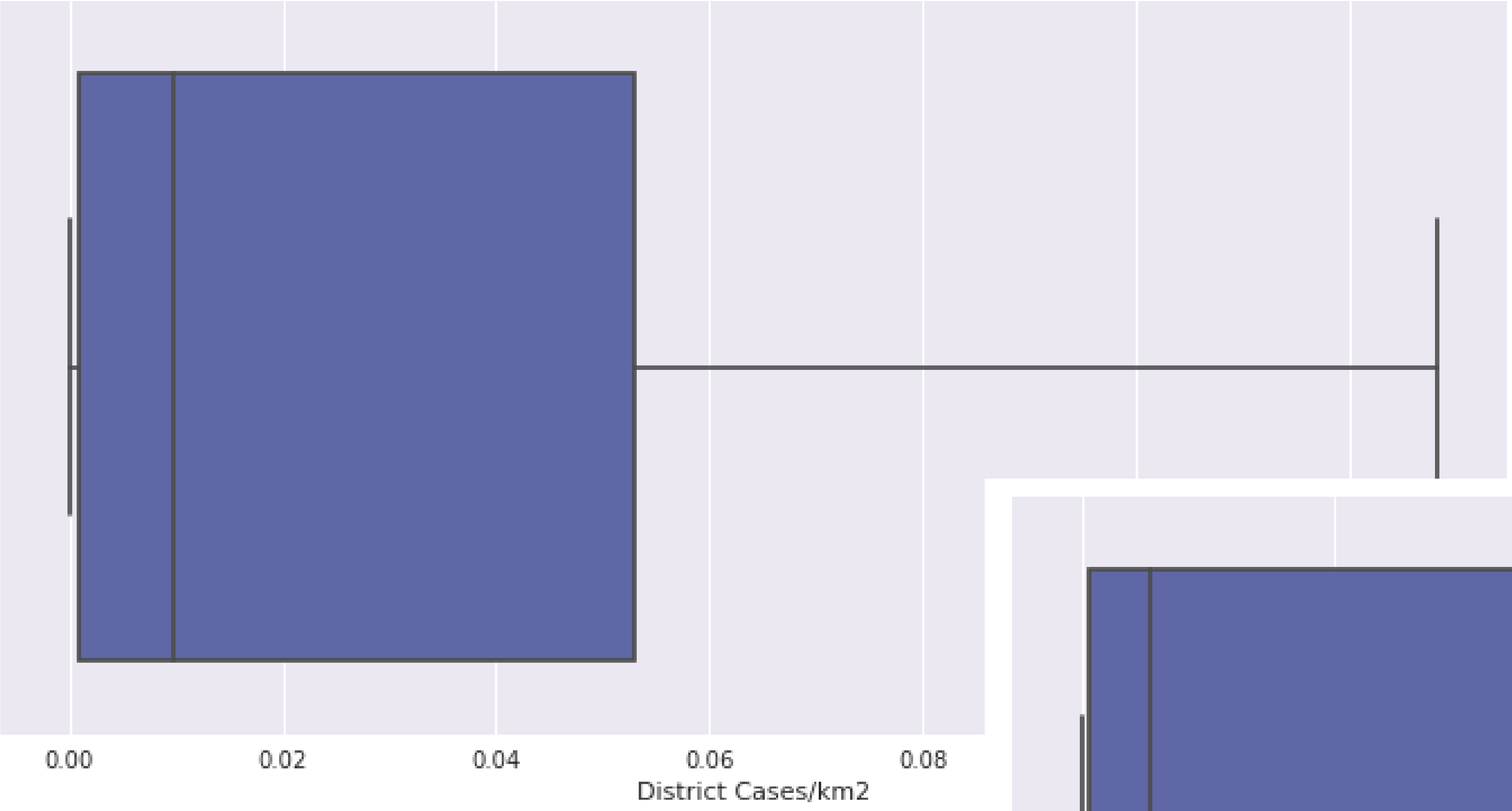




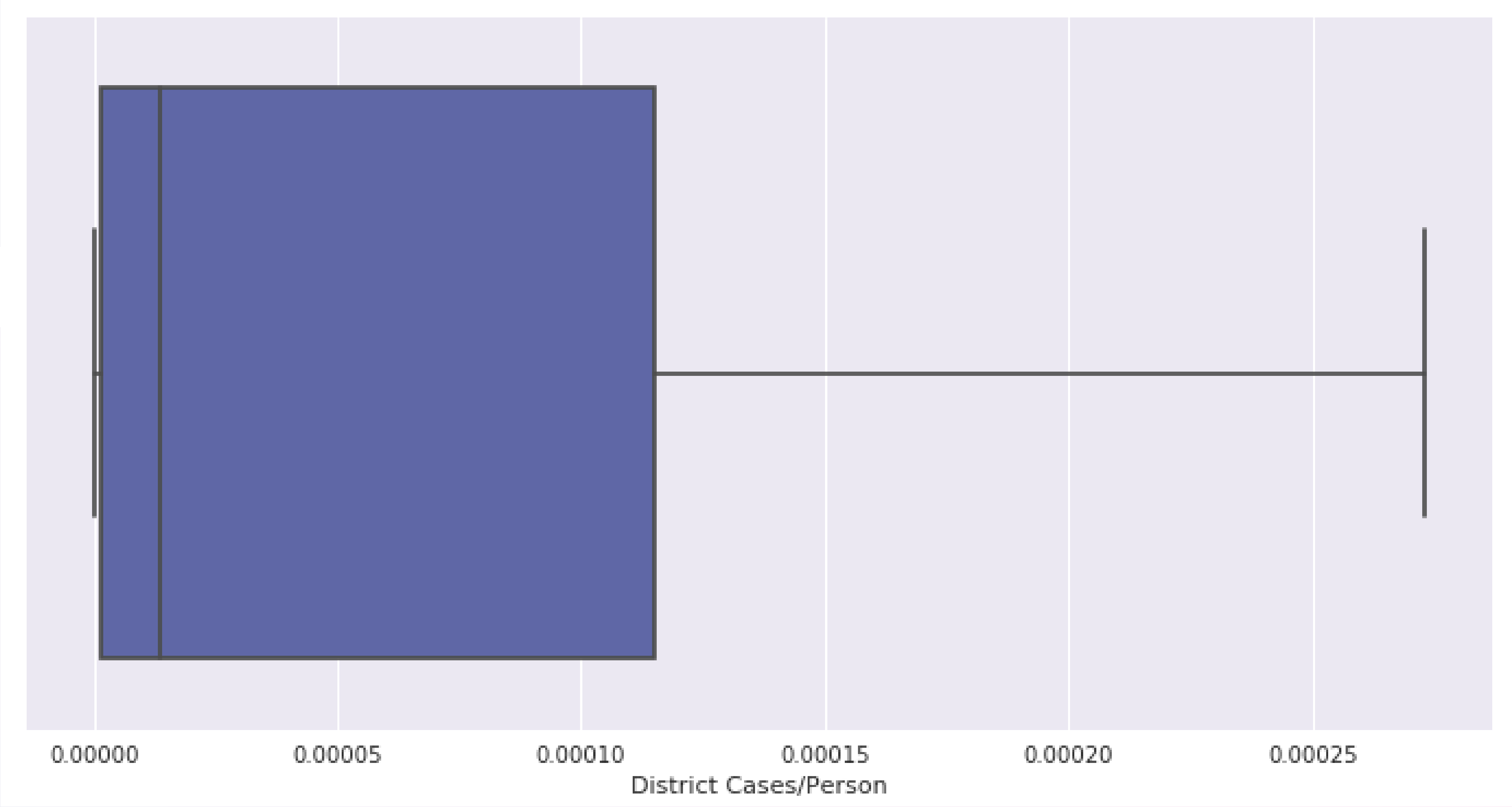
Histograms of Number of Ratings

Notice the highly extreme values, hence a use of a standard scaler would be needed.



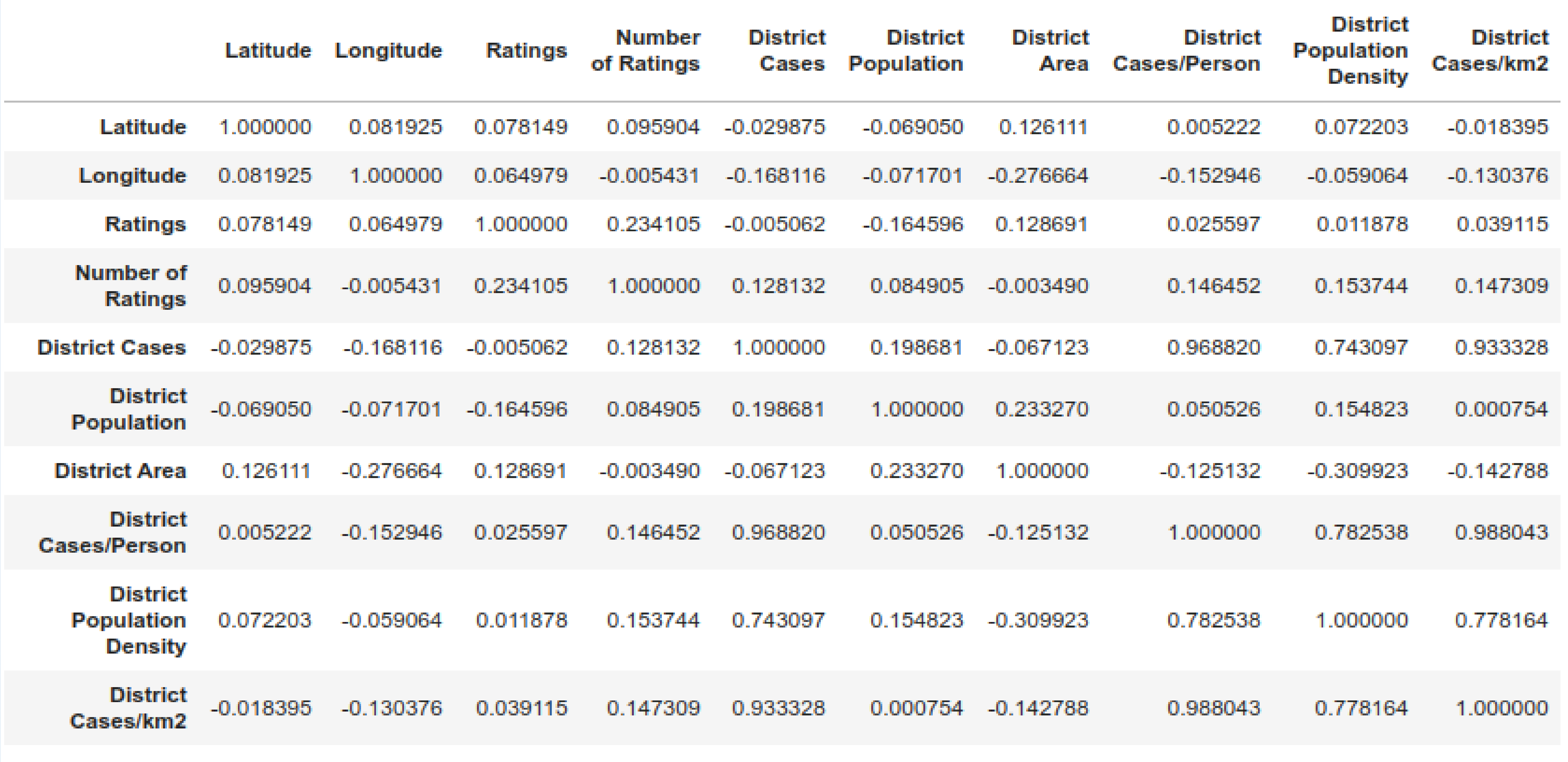


Boxplots of two possible features



Only one of these features is sufficient since they are extremely similar and highly correlated (see image on next slide)

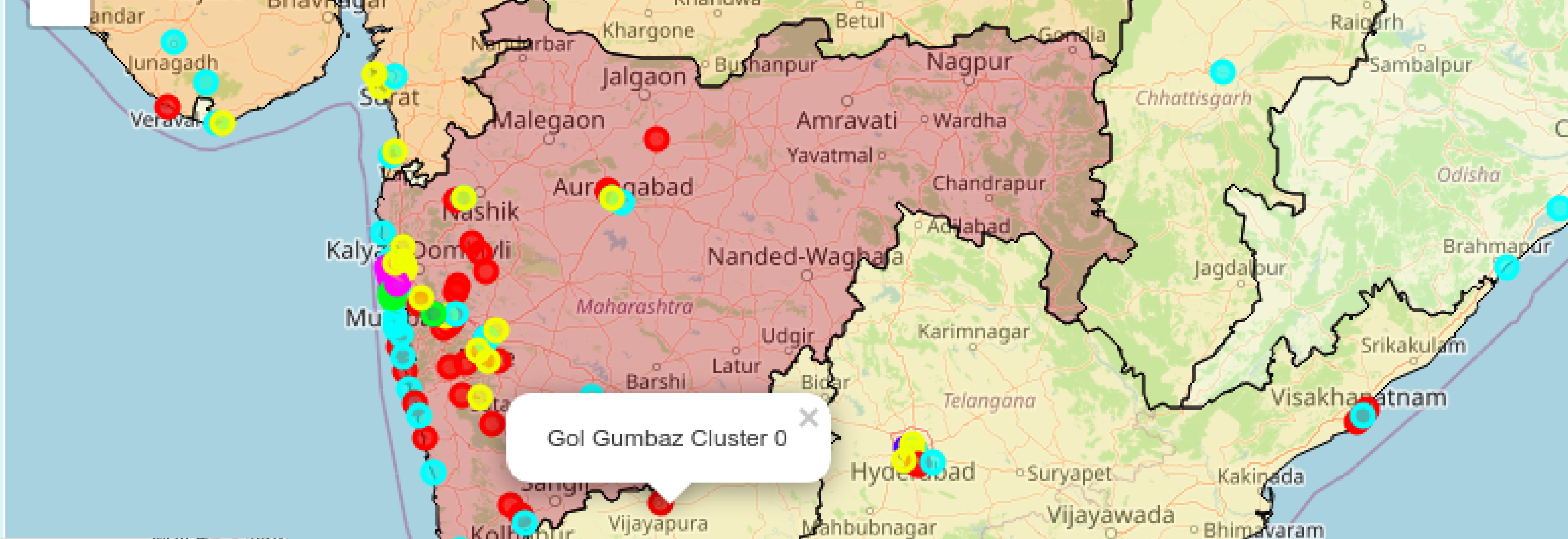
Correlation Coefficients

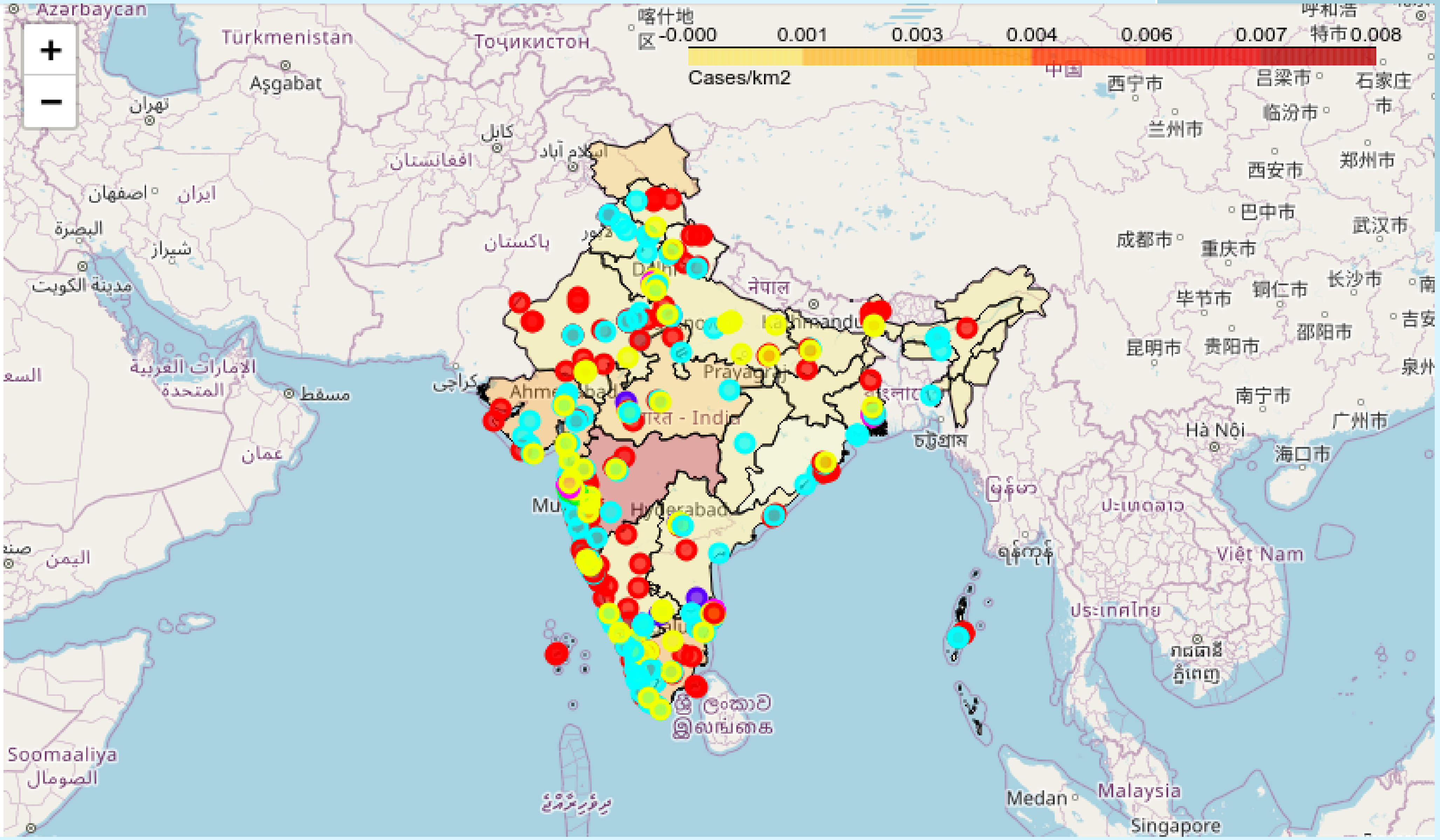


	Latitude	Longitude	Ratings	Number of Ratings	District Cases	District Population	District Area	District Cases/Person	District Population Density	District Cases/km2
Latitude	1.000000	0.081925	0.078149	0.095904	-0.029875	-0.069050	0.126111	0.005222	0.072203	-0.018395
Longitude	0.081925	1.000000	0.064979	-0.005431	-0.168116	-0.071701	-0.276664	-0.152946	-0.059064	-0.130376
Ratings	0.078149	0.064979	1.000000	0.234105	-0.005062	-0.164596	0.128691	0.025597	0.011878	0.039115
Number of Ratings	0.095904	-0.005431	0.234105	1.000000	0.128132	0.084905	-0.003490	0.146452	0.153744	0.147309
District Cases	-0.029875	-0.168116	-0.005062	0.128132	1.000000	0.198681	-0.067123	0.968820	0.743097	0.933328
District Population	-0.069050	-0.071701	-0.164596	0.084905	0.198681	1.000000	0.233270	0.050526	0.154823	0.000754
District Area	0.126111	-0.276664	0.128691	-0.003490	-0.067123	0.233270	1.000000	-0.125132	-0.309923	-0.142788
District Cases/Person	0.005222	-0.152946	0.025597	0.146452	0.968820	0.050526	-0.125132	1.000000	0.782538	0.988043
District Population Density	0.072203	-0.059064	0.011878	0.153744	0.743097	0.154823	-0.309923	0.782538	1.000000	0.778164
District Cases/km2	-0.018395	-0.130376	0.039115	0.147309	0.933328	0.000754	-0.142788	0.988043	0.778164	1.000000

Result

06 —





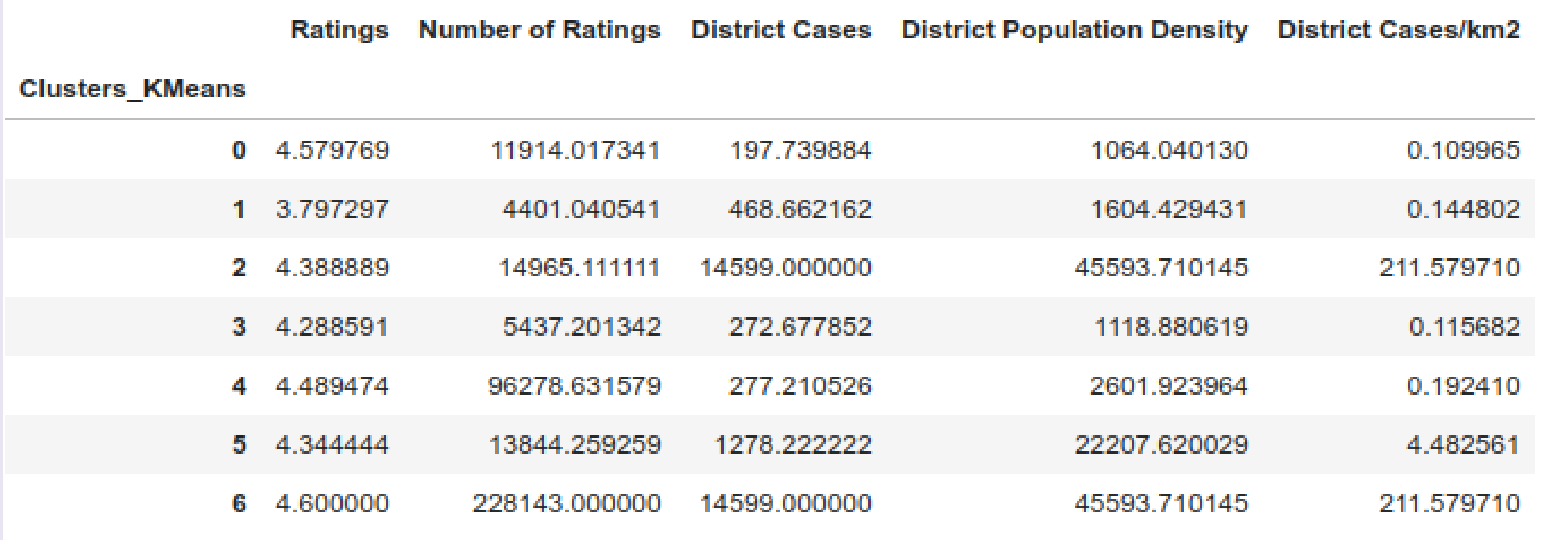
Algorithm used: K-Means Clustering

Number of clusters: 7 (combined two clusters later)

Used the choropleth map of the state coronavirus cases as the base

Discussion

07 —



Clusters_KMeans	Ratings	Number of Ratings	District Cases	District Population Density	District Cases/km2
0	4.579769	11914.017341	197.739884	1064.040130	0.109965
1	3.797297	4401.040541	468.662162	1604.429431	0.144802
2	4.388889	14965.111111	14599.000000	45593.710145	211.579710
3	4.288591	5437.201342	272.677852	1118.880619	0.115682
4	4.489474	96278.631579	277.210526	2601.923964	0.192410
5	4.344444	13844.259259	1278.222222	22207.620029	4.482561
6	4.600000	228143.000000	14599.000000	45593.710145	211.579710

Cl**uster 0** (**Red**) - Very high rating, moderately popular, least crowded and least number of cases - Highly recommended!
Cluster 1 (**Yellow**) - Very low rating, not popular, moderately crowded and moderate number of cases - Recommended for safety but not hugely for enjoyment.
Cluster 3 (**Light blue**) - Medium rating, not popular, not very crowded and medium number of cases - Recommended.
Cluster 4 (**Dark blue**) - High rating, highly popular, moderately crowded and moderate number of cases - Moderately recommended.
Cluster 5 (**Pink**) - Medium rating, moderately popular, highly crowded and high number of cases - Not recommended.
Cluster 2 (**Green**) - High rating, moderately popular, extremely crowded and extremely high number of cases - Avoid at all costs.

Discussion

07 —

Possible way forwards:

1. We can further use a district choropleth rather than one which looks at just states. This would help us understand which districts would be more suitable to travel through. This would take a little more time as updated GeoJSON files are not readily available and many districts are called by different names.
2. We can further divide tourist attractions based on type of destination to present it to stakeholders who care only about a certain type.
3. Further zoom in on a state (and perhaps neighbouring states), and conduct a similar deeper analysis for a particular state. This would seem like the next best thing to do considering that each state in India is massive in terms of touristic destinations.
4. Cost can be another feature to be included since it plays a vital role in deciding the right location to travel to.
5. APIs from tripadvisor, which has more wholesome travelling metrics could be used (Unfortunately, they do not allow its use for academia).
6. With coronavirus on the constant rise, it also gives us an opportunity to keep updating this project (can be easily done through the code in the notebook) as and when coronavirus numbers change continuously.
7. With better data, a density based clustering system can be incorporated.

Conclusion

08

This idea for the project was based on what the future of travelling could look like. With the world in a state of confusion, this project could serve as a data based solution/guide to those wanderlust travelholics who will now look towards domestic tourism.

Travelling to many is a way to understand themselves better and learn things you just cannot in a classroom. This project shows that options are available even if safety is our top priority, it is just a matter of finding those options.

Thank you!